Detecting Language Model Attacks with Perplexity

Year: 2023 | Citations: 340 | Authors: Gabriel Alon, Michael Kamfonas

---

Abstract

A novel hack involving Large Language Models (LLMs) has emerged, leveraging adversarial suffixes to trick models into generating perilous responses. This method has garnered considerable attention from reputable media outlets such as the New York Times and Wired, thereby influencing public perception regarding the security and safety of LLMs. In this study, we advocate the utilization of perplexity as one of the means to recognize such potential attacks. The underlying concept behind these hacks revolves around appending an unusually constructed string of text to a harmful query that would otherwise be blocked. This maneuver confuses the protective mechanisms and tricks the model into generating a forbidden response. Such scenarios could result in providing detailed instructions to a malicious user for constructing explosives or orchestrating a bank heist. Our investigation demonstrates the feasibility of employing perplexity, a prevalent natural language processing metric, to detect these adversarial tactics before generating a forbidden response. By evaluating the perplexity of queries with and without such adversarial suffixes using an open-source LLM, we discovered that nearly 90 percent were above a perplexity of 1000. This contrast underscores the efficacy of perplexity for detecting this type of exploit.